Talanta 83 (2011) 1260–1268

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Classification of jet fuels by fuzzy rule-building expert systems applied to three-way data by fast gas chromatography—fast scanning quadrupole ion trap mass spectrometry

Xiaobo Sun[a], Carolyn M. Zimmermann[a], Glen P. Jackson[a], Christopher E. Bunker[b], Peter B. Harrington[a,*]

[a] Center for Intelligent Chemical Instrumentation, Clippinger Laboratories, Department Of Chemistry and Biochemistry, Ohio University, Athens, OH 45701, USA
[b] Air Force Research Laboratory, Propulsion Directorate, Wright Patterson Air Force Base, Dayton, OH 45433, USA

ARTICLE INFO

Article history:
Available online 8 June 2010

Keywords:
Classification
Jet fuel
Fuzzy rule-building expert system
Gas chromatography
Fast scanning
Quadrupole ion trap mass spectrometry
Chemometrics

ABSTRACT

A fast method that can be used to classify unknown jet fuel types or detect possible property changes in jet fuel physical properties is of paramount interest to national defense and the airline industries. While fast gas chromatography (GC) has been used with conventional mass spectrometry (MS) to study jet fuels, fast GC was combined with fast scanning MS and used to classify jet fuels into lot numbers or origin for the first time by using fuzzy rule-building expert system (FuRES) classifiers. In the process of building classifiers, the data were pretreated with and without wavelet transformation and evaluated with respect to performance. Principal component transformation was used to compress the two-way data images prior to classification. Jet fuel samples were successfully classified with $99.8 \pm 0.5\%$ accuracy for both with and without wavelet compression. Ten bootstrapped Latin partitions were used to validate the generalized prediction accuracy. Optimized partial least squares (o-PLS) regression results were used as positively biased references for comparing the FuRES prediction results. The prediction results for the jet fuel samples obtained with these two methods were compared statistically. The projected difference resolution (PDR) method was also used to evaluate the fast GC and fast MS data. Two batches of aliquots of ten new samples were prepared and run independently 4 days apart to evaluate the robustness of the method. The only change in classification parameters was the use of polynomial retention time alignment to correct for drift that occurred during the 4-day span of the two collections. FuRES achieved perfect classifications for four models of uncompressed three-way data. This fast GC/fast MS method furnishes characteristics of high speed, accuracy, and robustness. This mode of measurement may be useful as a monitoring tool to track changes in the chemical composition of fuels that may also lead to property changes.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Research on the analysis of fuel is important and has been applied to safety assurance [1], workers' health protection [1,2], arson and forensic investigation [3–5] energy study [6,7] and environmental inspection [8–10]. To ensure aircraft fuel safety and quality requirements to be "clean" and "dry", classification of jet fuels is extremely important because quality degradation may occur as a result of aging, contamination, mislabeling, and even adulteration. Therefore, classification by lot number can help characterize and establish the provenance of fuels.

The analytical methodologies used to characterize jet fuels include gas chromatography coupled with mass spectrometry (GC/MS) [11–14], GC coupled with other detectors such as a flame ionization detector (GC-FID) [15], near-infrared (NIR) [16–19] and mid-infrared (mid-IR) [1,20,21], high performance liquid chromatography (HPLC) [13,22–24] and ^{13}C NMR spectroscopy [22,23,25].

Modern methods of analysis may yield overwhelming quantities of data so that usually only fractions of the acquired data are used in the decision making process. For example, many GC/MS studies rely on the total ion current (TIC) chromatograms to classify jet fuels. Chemometrics provides a framework to utilize all the information acquired during the measurement to solve complex problems such as classification of fuels. Chemometric data pretreatment methods commonly used in the study of fuels include spectral baseline-correction and retention time alignment [26–29], data compression [30], etc. Principal component analysis (PCA) is useful for dimension reduction in the study of jet fuels and other petroleum products [3,26]. Chemometric methods used for classification include artifi-

* Corresponding author.
E-mail address: peter.harrington@ohio.edu (P.B. Harrington).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.05.063

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE
Classification of jet fuels by fuzzy rule-building expert systems applied to three-way data by fast gas chromatography-fast scanning quadrupole ion trap mass spectrometry

5a. CONTRACT NUMBER
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER
5e. TASK NUMBER
5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory,Propulsion Directorate,Wright Patterson AFB,OH,45433

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)
11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A fast method that can be used to classify unknown jet fuel types or detect possible property changes in jet fuel physical properties is of paramount interest to national defense and the airline industries. While fast gas chromatography (GC) has been used with conventional mass spectrometry (MS) to study jet fuels, fast GC was combined with fast scanning MS and used to classify jet fuels into lot numbers or origin for the first time by using fuzzy rule-building expert system (FuRES) classifiers. In the process of building classifiers, the data were pretreated with and without wavelet transformation and evaluated with respect to performance. Principal component transformation was used to compress the two-way data images prior to classification. Jet fuel samples were successfully classified with 99.8?0.5% accuracy for both with and without wavelet compression. Ten bootstrapped Latin partitions were used to validate the generalized prediction accuracy. Optimized partial least squares (o-PLS) regression results were used as positively biased references for comparing the FuRES prediction results. The prediction results for the jet fuel samples obtained with these two methods were compared statistically. The projected difference resolution (PDR) method was also used to evaluate the fast GC and fast MS data. Two batches of aliquots of ten new samples were prepared and run independently 4 days apart to evaluate the robustness of the method. The only change in classification parameters was the use of polynomial retention time alignment to correct for drift that occurred during the 4-day span of the two collections. FuRES achieved perfect classifications for four models of uncompressed three-way data. This fast GC/fast MS method furnishes characteristics of high speed, accuracy, and robustness. This mode of measurement may be useful as a monitoring tool to track changes in the chemical composition of fuels that may also lead to property changes.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

cial neural networks (ANNs) [14], soft independent modeling class analogy (SIMCA) [5,14], K-Nearest neighbor (KNN) [14,31], partial least squares (PLS) regression [17,28], multivariate least squares (MLS) regression, linear discriminant analysis (LDA) [32], multiple linear regression (MLR) [33,34], etc. The fuzzy rule-building expert system (FuRES) [35] has demonstrated utility for classification of jet fuels because the models are reproducible and amenable to interpretation [27].

As a major instrumental method, GC has found extensive application in research of jet fuels because of the volatile nature of jet fuel components and the powerful separation capacity of GC. Especially when GC is coupled with MS, it can give a three-way data set (GC, MS and numbers of samples), which is rich in information of the composition of fuel products. Two-dimensional gas chromatography (GC × GC) has been also widely used because of the advantages of larger peak capacity, higher separation capacity, an increase in sensitivity and a faster analysis speed [36,37].

Doble et al. conducted a GC/MS study to classify premium and regular gasoline and their seasonal formulation (winter or summer) [3]. By using the Mahalanobis distances calculated from the principal components scores, a classification rate of about 80–93% was achieved over the premium and regular gasoline samples, but only a 48–62% classification rate was obtained for the winter and summer samples as the sub-groups. When an ANN model, which was trained by back propagation and conjugate gradient algorithms, was used, a 100% classification rate for the premium and regular samples and a 96% classification rate for the summer and winter sub-groups were achieved. Although this method was applied to the gasoline products instead of jet fuels, it also can be used in the prediction of jet fuel properties.

In the extensive study about jet fuels conducted by the Naval Research Lab (NRL), jet fuel property prediction was the focus, in which they have been successful [17,38–40]. They used near-infrared (NIR) spectroscopy, Raman spectroscopy and GC with a flame ionization detector (GC-FID) or a mass spectrometer detector (GC-MS) for data collection. The chemometric technique used by these researchers focused on partial least squares (PLS) regression.

According to the definition given by Matisová et al., fast GC has a separation time per sample of a few minutes and a speed enhancement factor of 5–30 while the plate number can be kept comparable to the conventional gas chromatography [41]. By using microbore columns, faster temperature programming, a faster carrier gas speed, and a higher head pressure, etc., fast GC separation can be realized. Compared with conventional GC, fast GC offers the advantages of a tremendous improvement in laboratory throughput, a much lower cost per sample, a shorter time needed for the analysis, and especially the possibility of the usage as an online monitoring means. However, fast GC suffers from limited chromatographic resolution especially when it is used to separate very complicated samples such as jet fuels. It also requires a faster detection method to match the faster separation speed.

As one of the most common, sensitive, and informative detectors for GC, MS has promise for the composition-property correlation study of jet fuels. Time-of-flight (ToF) mass spectrometers are capable of very fast data acquisition rates (kHz duty cycles are possible.) as summarized in a review about application of GC/ToF-MS [42]. However, they have the disadvantage of relatively higher costs to purchase and maintain compared to ion trap mass spectrometers.

Most conventional scanning MS methods, such as the quadrupole ion trap MS, have limited scan speeds of 5500 Th/s (scan rate parameter: 0.18 ms/Th), or acquisition rate of ~3 Hz, which will be insufficient when it is used as a detection means for fast GC. Yang and Bier proposed a fast ion trap MS scan strategy with a scan rate as fast as 66 660 Th/s (scan rate parameter: 0.015 ms/Th), which is 12 times the scan rate of conventional MS [43]. Yang and Bier noted a fortuitous and unique result of increasing the scan rate of the quadrupole ion trap, an overall signal-to-noise improvement through the reduction in space charge effects and a decrease in peak widths (hence an increase in peak heights). Fast scanning in QITs has the disadvantages of decreased mass resolution and a possible decrease in mass accuracy. The primary advantage is that fast scanning QITs are better suited for coupling with time-limited fast chromatographic separations that yield peaks in narrow time windows.

FuRES is a pattern recognition technique devised by Harrington [35]. It has been successfully used in classification of complex data sets [27,44,45], especially jet fuel data [27]. FuRES provides an easy-to-understand mechanism of inference that is represented as a classification tree. The principal component transformation (PCT) is used to reduce the computational load of the FuRES nonlinear optimization that is required to construct the multivariate rules. FuRES is a robust and efficient pattern recognition method.

In the present work, for the first time fast GC coupled with a fast scanning quadrupole ion trap mass spectrometer was applied as the three-way data collection method [46]. The data were imported and compressed by using principal component transformation before being subjected to the FuRES and PLS classification. As a comparison, the same data set was also compressed by both wavelet transformation and principal component transformation before they were subjected to the FuRES and PLS classification. A classification accuracy of 99.8 ± 0.5% was obtained using the FuRES classifier and fast chromatography with fast scanning mass spectrometry for both with wavelet compression and without wavelet compression.

The work conducted by the NRL focused on mainly the individual property prediction of jet fuel samples to screen possible property changes [17,38–40]. Different from work done by the NRL, this work emphasized on whole sample information derived from the three-way fast GC/fast MS data set as the foundation of classification. If the fuel is recognized by its lot number, then the physical properties can be deduced to be similar as those of the lot. The accuracy, high speed, and reliability of this method demonstrate the great potential for screening jet fuels.

2. Experimental section

2.1. Reagents and sample preparation

The jet fuel samples were provided by the Air Force Research Laboratory of Wright Patterson Air Force Base (Dayton, OH). All the samples were stored in borosilicate glass vials at room temperature and used as received.

Twenty samples were chosen randomly from a library of 200 samples, which would ensure the sampling probability of all samples to be equal, and diluted 1:50 with pentane of HPLC grade (Sigma Aldrich). Dilution of samples with pentane was necessary to avoid detector saturation. Because the sampling was random from the library, four Jet A samples, twelve JP-8 samples, two JPTS samples, one JP-8+100 sample and one Jet A-1 sample were chosen for the experiments, which were representative of the distribution of fuels in the library. All the samples were freshly prepared and measured following a random block design with time as the blocking factor. The jet fuel types, sample IDs and available properties are given in Tables 1 and 2.

2.2. Instrumentation and methods

Five replicates were run for each sample by following an autosampler sequence generated by random block design. A solvent blank was run before and after each block to validate the lack of carryover with three cycles of syringe washes before and after the injection. All experimental data were collected on a Trace-GC

Table 1
Types, IDs and available properties of the samples used in the comparison of the FF-CI mode with the FN-CI mode.

	Sample ID	3528	3998	3752	3488	2851	2829	2882	2885	4198	2993
	Fuel type	JP-8	Jet A	JP-8	JP-8	JP-8	JP-8	JP-8	JP-8	JP-8	Jet A
D3241	Tube deposit rating, visual	1	1	1	1	<1	1	<1	<1	<2	1
D3241	Change in pressure (mm Hg)	2	0	0	0	0	0	1	0	3	n/a
D5972	Freezing point (°C) (automatic)	−50	−40	−50	−49	−49	−47	−49	−51	−50	−59
D86	IBP (°C)	171	174	132	156	160	157	148	145	179	n/a
D86	10% recovered (°C)	189	192	160	182	182	177	175	170	189	n/a
D86	20% recovered (°C)	195	199	168	190	188	185	183	180	193	n/a
D86	50% recovered (°C)	210	220	190	209	206	203	203	204	206	n/a
D86	90% recovered (°C)	239	259	234	240	242	235	237	246	234	n/a
D86	EP (°C)	257	278	254	259	268	259	260	269	250	n/a
D86	Residue (vol%)	1.3	1.5	1.3	1	1.2	1.3	1.3	1.3	1.3	n/a
D86	Loss (vol%)	0.5	0.2	0.8	0.7	0.5	0.2	0.9	0.6	0.2	n/a
D381	Existent gum (mg/100 mL)	4	2.4	0.4	1.4	1.2	3.2	1	4	0.4	n/a
D93	Flash point (°C)	58	60	41	52	52	49	47	48	64	−54
SPEC\F	Filtration time (min)	6	9	5	6	4	4	4	4	8	n/a
D5452	Particulate (mg/L matter)	0.4	0.6	0.2	0.4	0.4	0.5	0.4	0.4	0.4	n/a

2000 gas chromatograph (GC) equipped with a Thermo Finnigan Polaris Q quadrupole ion trap mass spectrometer (MS) (Thermo Electron Corporation, San Francisco, CA, USA) as the detector. The gas chromatograph was also equipped with a TRIPLUS AS autosampler (Thermo Scientific). The Xcalibur software version 1.4 (Thermo Scientific) was used for the instrument control and data collection. Enhanced scan rates were used through modification of the custom tune program through the freely available XDK command package (Thermo Scientific) using Visual Basic 6.0 (Microsoft Corporation, Redmond, Washington, USA). Because of the limited acquisition rate of the analogue to digital converter, there is a trade-off between scan rate parameter and sampling points per Th: faster scanning results in fewer data points per unit Th and therefore poorer mass resolving power. The scan rate parameter of 0.06 ms/Th offered the best balance among speed, signal-to-noise ratio improvement, and mass resolution for the present study.

With an initial set of ten samples (see Table 1) two experimental modes were evaluated: fast GC separation with fast MS scan (the FF mode), and fast GC with conventional MS scan (the FN mode) with the latter as a reference method (FF-CI and FN-CI; both modes used chemical ionization (CI)).

A second set of ten samples (see Table 2) was analyzed using the FF-CI mode for the purpose of validating the procedure. One batch of aliquots was collected from this second set of samples and was analyzed and a second batch of aliquots was collected and was analyzed 4 days later. Each set of aliquots was prepared independently.

CI operated under the positive ion mode with isobutane (99.00%, Airgas) as the reagent gas at a flow rate of 0.6 mL/min. The mass scan range was from 60.00 to 425.00 Th for both MS configurations. The fast GC-normal scan MS was selected as a reference method because fast GC has already been demonstrated in the literature for fuel samples [29].

The separation was accomplished with a 0.2 μm film of polydimethyldiphenyl siloxane (5% phenyl) [DB-5, Agilent Technologies] wall coated open tubular column with a 5.0 m length and a 0.10 mm internal diameter. The initial temperature was 50 °C and held for 1 min, increased at a rate of 30 °C/min to 220 °C, and held for 1 min at 220 °C. A 0.3 min solvent delay was used under the split mode with a split ratio of 1:20. A flow rate of 1.5 mL/min of carrier gas helium was maintained by the flow controller. The conditions of GC and MS for the FF-CI mode are summarized in Table 3.

2.3. Data processing

2.3.1. General information

The data collected by the Xcalibur software version 1.4 (Thermo Scientific) were imported into and processed with the MATLAB version R2010a software (The MathWorks Inc., Natick, MA) on a home-built computer equipped with an Intel Core i7 940 processor with 12 GB of DDR3 RAM. The operating system was Microsoft Windows XP x64 Professional SP1.

2.3.2. Data compression

For the purpose of comparison, the data set collected from samples given in Table 1 was treated with and without wavelet compression. For the two-dimensional wavelet compression both retention time (RT) and mass-to-charge ratio dimensions were

Table 2
Types, IDs and available properties of the samples used in the prediction of unknown samples under the FF-CI mode.

	Sample ID	4131	3869	3520	3517	3737	4160	4195	4255	4773	4188
	Fuel type	JP-8	JPTS	JP-8	Jet A	JP-8+100	JP-8	JPTS	Jet A-1	Jet A	JP-8
D3241	Tube deposit rating, visual	1	n/a	1	1	<1	<1	n/a	<1	1	4
D3241	Change in pressure (mm Hg)	0	1	0	1	1	1	1	1	1	12
D5972	Freezing point (°C) (automatic)	−54	−56	−57	−52	−56	−49	−60	−54	−47	−49
D86	IBP (°C)	155	157	147	n/a	148	153	160	n/a	n/a	162
D86	10% recovered (°C)	171	165	170	n/a	164	177	167	n/a	n/a	181
D86	20% recovered (°C)	180	n/a	176	n/a	170	182	n/a	n/a	n/a	188
D86	50% recovered (°C)	204	179	192	n/a	196	200	179	n/a	n/a	206
D86	90% recovered (°C)	245	220	227	n/a	243	237	215	n/a	n/a	241
D86	EP (°C)	267	241	253	n/a	266	260	241	n/a	n/a	271
D86	Residue (vol%)	1.4	1.0	1.2	n/a	1.4	1.3	1.0	n/a	n/a	1.4
D86	Loss (vol%)	0.7	0.6	0.6	n/a	0.3	0.2	0.2	n/a	n/a	0.1
D381	Existent gum (mg/100 mL)	1.6	0.6	0.2	n/a	1.2	1.8	0.8	n/a	n/a	20.0
D93	Flash point (°C)	48	47	45	54	44	51	48	55	49	52
SPEC\F	Filtration time (min)	5	n/a	5	n/a	7	7	n/a	n/a	n/a	7
D5452	Particulate (mg/L matter)	0.8	0.3	0.5	n/a	0.4	0.6	0.2	n/a	n/a	0.9

Table 3
GC separation and MS scan conditions for the FF-CI mode.

Column	DB-5: 5.0 m × 0.10 mm × 0.2 μm
Temperature program	Initial 50 °C held for 1 min Ramp 30 °C/min Final 220 °C held for 1 min
Instrument analysis time	7.7 min
Injector temperature	250 °C
Transfer line temperature	280 °C
Carrier gas	Helium, 1.5 mL/min
Injection mode	Injection volume of 1 μL; split ratio of 20
Scan rate parameter	0.06 ms/Th
Sampling points (SAMP)	11.3/Th
Solvent delay	0.3 min
Mass range	60–425 Th
Ion source temperature	200 °C
Reagent gas	Isobutane
Flow rate of reagent gas	0.6 mL/min

compressed. The data were compressed using Villasenor biorthogonal wavelets: first the mass spectral dimension was compressed and then the RT dimension was compressed to yield a compression to 1/16 of the original size. The RT and mass scales were fixed to a constant value among the different samples by binning. The resulting point spacing for the mass and the RT orders were 0.1 Th and 0.01 min, respectively, after the wavelet compression. Each two-way data object was normalized to unit vector length.

2.3.3. Principal component analysis

Principal component analysis was used to visualize the clustering of the object scores for both the jet fuel properties and the GC/MS data. The data were centered by subtracting the average of the data objects from each object in the data set prior to calculating the principal components by singular value decomposition.

2.3.4. Projected difference resolution (PDR) metric [27]

For complex data sets, the distribution of objects and classes cannot be accurately assessed by looking at the principal component scores, especially when the variance spanned by the first two components is less than 90% as is often the case. Visually assessing 3D plots of principal component scores is always a bad idea because it is not quantitative. As a powerful tool, the PDR quantitative metric was devised that represents separations of clusters in a multidimensional space in the context of chromatographic resolution [27]. The difference vector between two class means of the objects is calculated first. Then the objects of the two classes are projected onto the difference vector to yield a scalar set of scores. The projections are used similar to those in the standard chromatographic resolution equation. The stepwise calculations follow. First, the difference vector between two class means is calculated.

$$\mathbf{d_{a,b}} = \bar{\mathbf{x}}_\mathbf{a} - \bar{\mathbf{x}}_\mathbf{b} \quad (1)$$

for which $\bar{\mathbf{x}}_\mathbf{a}$ and $\bar{\mathbf{x}}_\mathbf{b}$ are the class means and $\mathbf{d_{a,b}}$ is the difference vector between $\bar{\mathbf{x}}_\mathbf{b}$ and $\bar{\mathbf{x}}_\mathbf{a}$. Objects are row vectors.

$$p_i = \mathbf{x_i d^T} \quad (2)$$

for which p_i is the inner product of data object $\mathbf{x_i}$ and the class difference vector $\mathbf{d}$. The resolution of two classes then can be calculated according to the equation below

$$R_s = \frac{\left|\bar{p}_a - \bar{p}_b\right|}{2 \times (s_a + s_b)} \quad (3)$$

for which p_a and p_b are differences of the averages of the projections; s_a and s_b are the standard deviations of the two classes. As with chromatographic resolution, when the R_s value is larger than 1.5, the classes are considered resolved. In this work, the PDR method was used to evaluate and optimize the data pretreatment steps.

2.3.5. FuRES and o-PLS classifiers

The class designees were binary encoded. FuRES does not have an adjustable parameter, such as the component number in PLS, so it does not require a separate set of data to optimize the model. For the o-PLS (in-house program) [46] model, the full set of latent variables was calculated. During prediction, the number of latent variables that yielded the lowest prediction error as defined by the predicted residual error sum of squares (PRESS) of each prediction data set was used to generate the best possible predictions. o-PLS acts as a positively biased reference method and if an equivalent or better FuRES prediction result is obtained, the FuRES method is validated.

Instead of using a single prediction set and a single model, three Latin partitions and ten bootstraps were used to provide a generalized validation of the classifiers. The results of the three prediction sets from each partition were pooled so that every object was used one time for prediction and twice for model building. The results were also used for two-way analysis of variance (ANOVA) comparisons between the FuRES and the o-PLS predictions. The prediction results were averaged across the 10 bootstraps to provide 95% confidence intervals.

2.3.6. Prediction of a novel set of samples

A set of ten new samples was randomly selected from the pool of 200 jet fuels samples. The two sets of samples were run 4 days apart and designated as batch A for the earlier collection and batch B for the later collection. These samples were analyzed independently including the dilution step. Batch A was used for model building and batch B was used for the prediction. Then the roles of the two sets of 50 objects were reversed with batch B used for model building and the batch A used for prediction.

Because RT drift occurred during the 4-day period separating the data collections retention time alignment was implemented. The three-way alignment is a standard procedure in our lab and was used without optimization (i.e., the default parameters of a single iteration and a third order polynomial was used). The average two-way image of the GC/MS data is calculated and used as a target. Each two-way image is aligned by using a third order polynomial to adjust the retention time to maximize the correlation coefficient of the two-way data object and the two-way average. The intensities are adjusted using linear interpolation (i.e., interp1 function in MATLAB). For prediction data, each prediction object was aligned to the two-way mean of the unaligned calibration data.

3. Results and discussion

3.1. Sample property analysis by hierarchical cluster analysis and PCA

Because some properties of the sample sets were not available, only nine samples were used in this assessment. Properties with units of temperature were evaluated so that distances in the dendrogram and among the PCA scores are differences in temperature and can be easily assessed. The hierarchical cluster analysis was conducted by calculating the average linkage distance [47]. The dendrogram obtained from the freezing point, initial boiling point, 10% recovered boiling point, 20% recovered boiling point, 50% recovered boiling point, 90% recovered boiling point, and end boiling point of samples 3528, 3998, 3752, 3488, 2851, 2829, 2882, 2885, and 4198 is given in Fig. 1. The principal component scores of the temperatures of the same sample properties are given in Fig. 2.

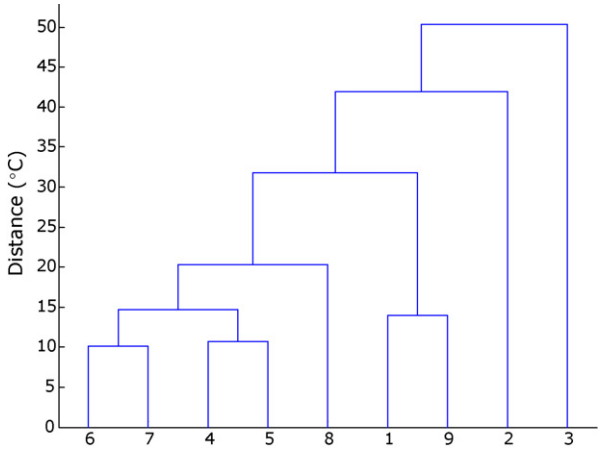

Fig. 1. Dendrogram of selected properties of samples 3528, 3998, 3752, 3488, 2851, 2829, 2882, 2885 and 4198.

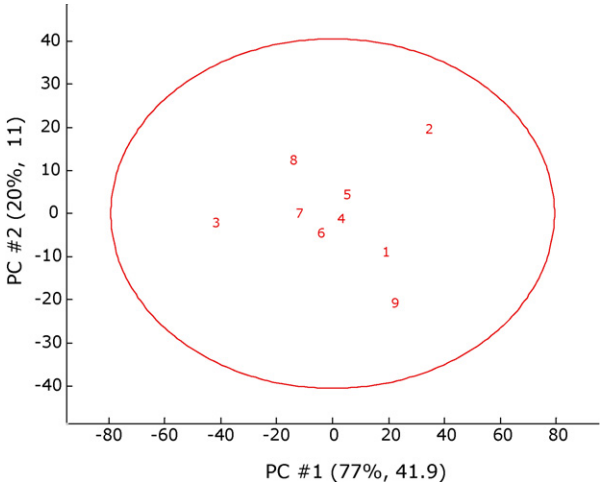

Fig. 2. PCA score plotting of selected properties of samples 3528, 3998, 3752, 3488, 2851, 2829, 2882, 2885 and 4198.

From both analyses there are differences and similarities among the properties of fuels with different lot numbers.

3.2. Instrumentation and GC/MS data

Using pentane and a solvent delay of 0.3 min may pose a limitation in that the most volatile components that elute during the 0.3 min solvent delay are not characterized. The solvent delay in our work was not relevant because the MS scanned a range of 60–425 Th, thus ions below 60 Th generated from the early eluting compounds would not be detected.

Five replicates for the 10 samples were collected using a random block design implemented by the autosampler to yield 50 GC/MS data objects. Each run took 7.7 min, which was one quarter of the conventional GC/MS separation time. As an example, the total ion current (TIC) chromatogram and average mass spectrum of sample 2885 are given in Fig. 3. With a short separation time, the chromatographic peaks were significantly overlapped so that it would be difficult to classify the jet fuel samples visually. However, by representing the data object as a two-way image (see Fig. 4), better resolution is apparent in the two-way data image although some peaks are not completely resolved.

3.3. Data compression

Wavelet transformation can provide compression of the analytical data. As an example, the original size of a two-way data object of sample 3998 was 4124×1349. After biorthogonal Villasenor wavelet (the Wavelab toolbox) compression was applied to the RT and mass-to-charge ratio orders, the data size was reduced to 1031×338, which is one sixteenth of the original size. This compression offers the advantages of smaller data size, which will result in a much shorter computing time while preserving the peaks and concomitantly improves the signal-to-noise ratio by removing high frequency noise components. The advantage of the biorthogonal wavelet compression is that peaks in the compressed data are not shifted in their location with respect to the retention time and mass-to-charge ratio orders. However, wavelet compression may result in signal loss, influencing results such as prediction rates as discussed in the following sections.

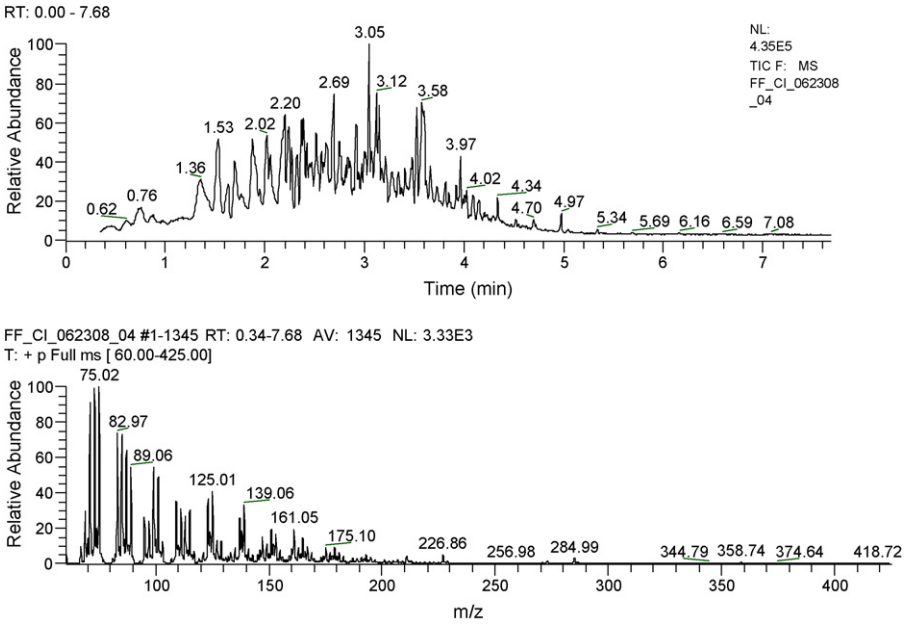

Fig. 3. TIC and average mass spectrum of sample 2885 under the FF-CI mode.

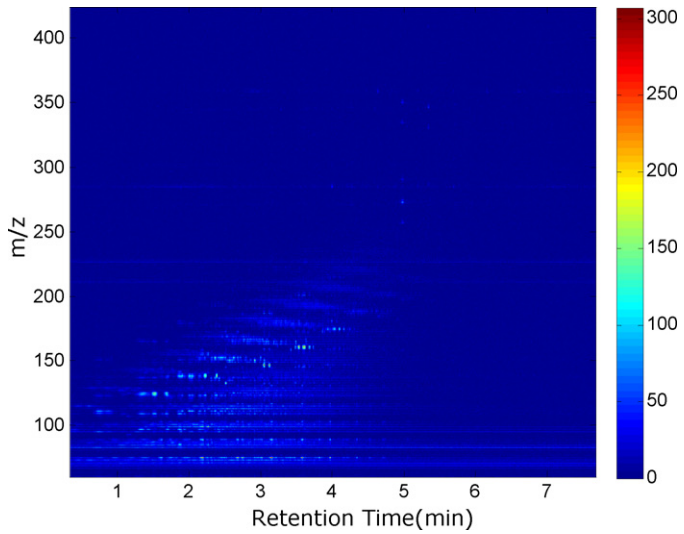

Fig. 4. Two-way data image of sample 2885 under the FF-CI mode reconstructed with MATLAB.

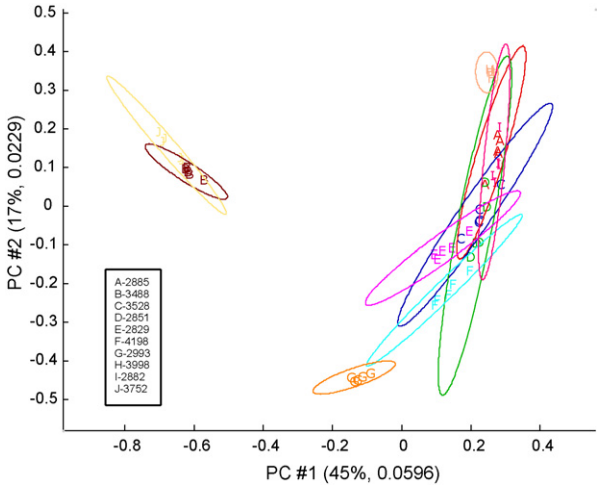

Fig. 5. The principal component score plotting of data collected with the FF-CI mode without wavelet compression.

3.4. PCA results

PCA scores of the data collected under the FF-CI mode were used to evaluate the clustering of the samples and replicates. The plot of scores of the two-way objects is given in Fig. 5. The first and the second principal components (PCs) account for 62% of the total variance. The ellipses are the 95% confidence interval around the means of each sample. The numbers in the parentheses give the relative and the absolute variance of the PCs. From the score plot, it can be concluded that some types of the jet fuel sample clusters are overlapped when projected onto the two PCs; however the clusters may be resolved in the higher dimensional data space.

3.5. PDR results

Analogous to chromatographic resolution, a PDR resolution of 1.5 is considered baseline separation of two clusters of data in the multidimensional data space. The minimum average resolutions of the jet fuel data with 10 bootstraps with and without wavelet compression were 2.3 ± 0.4 and 1.9 ± 0.1, respectively. Because these values were larger than 1.5, the class boundaries were completely separated in the multivariate data space. When the confidence intervals are considered, there was not a significant difference between these two resolutions. The geometric mean resolutions with and without wavelet compression respectfully were 6.6 ± 0.1 and 8.3 ± 0.5 and did differ significantly. Values larger than 1.5 indicate that successful classification should be achievable and some loss of signal is observed with the two-way wavelet compression.

3.6. FuRES and o-PLS classification results

Three Latin partitions with ten bootstraps were used to build 30 FuRES and o-PLS models from randomly selected subsets of the two-way data objects. For each bootstrap, the data were split into training and prediction sets by Latin partitions so that each spectrum was used only once in the prediction set and the same class distributions were maintained between training and prediction sets. Prior to constructing the classifiers the model-building data set was compressed using the PCT so that the size was further reduced to 34×34 or 33×33. There were two compressed sizes because 50 is not a multiple of 3, so the size of the model-building data set and prediction set varied by unity among the three partitions. The prediction data set was compressed by projection onto the same principal components that were calculated from the training data set.

Two-way ANOVA with interaction was used to compare results between the o-PLS control method and FuRES. The total run time on the computer was between 50 and 60 min for each evaluation that would construct 30 FuRES models and 30 o-PLS models. The three-way data set with and without wavelet compression was used to construct the FuRES and o-PLS classification models and to validate them with bootstrapped Latin partitions.

PDR values and the average prediction rates are reported with 95% confidence intervals for the ten bootstrapped Latin partitions with and without wavelet compression in Table 4. The minimum resolution measures the relative separation of the most overlapped pair of classes or fuel lots in the multivariate data space. The geometric mean of the PDR values gives an overall measure of the separation of all the combinations of pairs of classes.

For the FN-CI mode, the minimum resolution was 1.7 ± 0.7 before wavelet compression and 0.8 ± 0.2 after wavelet compression. The PDR measure reveals that compression deleteriously affected at least one pair of jet fuel lots. The geometric PDR

Table 4
Prediction and resolution results of FF-CI and FN-CI with and without wavelet compression.

	FF-CI		FN-CI	
	Uncompressed	WL compressed	Uncompressed	WL compressed
FuRES prediction rate (%)	99.8 ± 0.5	99.8 ± 0.5	97.4 ± 1.0	88 ± 1
o-PLS prediction rate (%)	100	100	97.8 ± 0.5	93 ± 2
Mean of minimum resolution	1.9 ± 0.1	2.3 ± 0.4	1.7 ± 0.7	0.8 ± 0.2
Geometric mean of minimum resolution	8.3 ± 0.5	6.6 ± 0.1	6.0 ± 0.5	2.9 ± 0.1

Note: FF-CI indicates the fast chromatography and fast MS scan under CI ionization; FN-CI indicates the fast chromatography and normal MS scan under CI ionization; WL designates a 2×2 biorthogonal wavelet compression to 1/16th of the data set size. Precision measures for classification and PDR are 95% confidence intervals calculated from ten bootstraps.

means were 6.0 ± 0.5 before wavelet compression and 2.9 ± 0.1 after the wavelet compression, which also indicates that for this case significant amounts of information were lost during wavelet compression. For the normal MS scan (FN-CI) mode a significant decrease in PDR is observed because there are fewer data points along the fast chromatographic order and the data are overcompressed with respect to retention time.

For the FF-CI mode, although the minimum PDR values of 1.9 ± 0.1 and 2.3 ± 0.4, before and after compression respectively, did not demonstrate a big difference, the geometric means of the projected difference resolutions respectively decreased from 8.3 ± 0.5 to 6.6 ± 0.1 for before and after compression. The decrease in PDR indicates characteristic information is lost after the compression for clusters of class objects that are well separated, but not enough to cause class overlap and accurate classification should be expected. Wavelet compression may improve signal-to-noise ratios, which may explain the marginal improvement in the minimum resolution.

For the normal MS scan this trend in PDR is also supported by the lower average prediction rates for FuRES and o-PLS classifications than those when the data set was uncompressed. For the FN-CI data, classification rates of FuRES and o-PLS were 88 ± 1% and 93 ± 2% with wavelet compression and 97 ± 1 and 97.8 ± 0.5 without wavelet compression. This substantial decrease in classification rate caused by wavelet compression indicates that with a fast separation, the conventional MS scan rate does not adequately resolve rapidly eluting compounds. The insufficient sampling with respect to retention time results in a further loss of characteristic information if the data are compressed as manifested by the worsened projected difference resolutions and prediction accuracies.

Alternatively, for the fast MS scan, wavelet compression did not exert an effect on the prediction rates for the FuRES classification (FF-CI with and without wavelet compression were both 99.8 ± 0.5%) and the o-PLS classification of 100% accuracy for both with and without wavelet compression. When fast MS scan was implemented, more data points were available with respect to the retention time. Although wavelet compression was used, the remaining data points still retained enough information to permit accurate classification of the samples. Therefore, fast scan MS is advantageous and necessary for fast GC separation for the purpose of classification of these jet fuel samples.

The FuRES classification tree with wavelet compression for the FF-CI mode is given in Fig. 6. In the classification tree, *H* represents the classification entropy. The numbers refer to the rule used for classification at each branch of the tree. Moving from the root to the leaves of the tree, the value of entropy decreases because the number of multiple classes at each rule is decreased. *Nc* gives the number of objects in each class. The FuRES classification tree has the most efficient classification structure with nine rules which are the minimum possible to resolve ten fuel samples.

o-PLS was used as a reference method in this work because it is positively biased with the latent variables optimized to give the lowest prediction error for each prediction set. The results of the FF-CI mode with and without wavelet compression both achieved 99.8 ± 0.5% for FuRES and 100% for the o-PLS. This result demonstrates the robustness of FuRES classifiers that performed as well as the positively biased reference method.

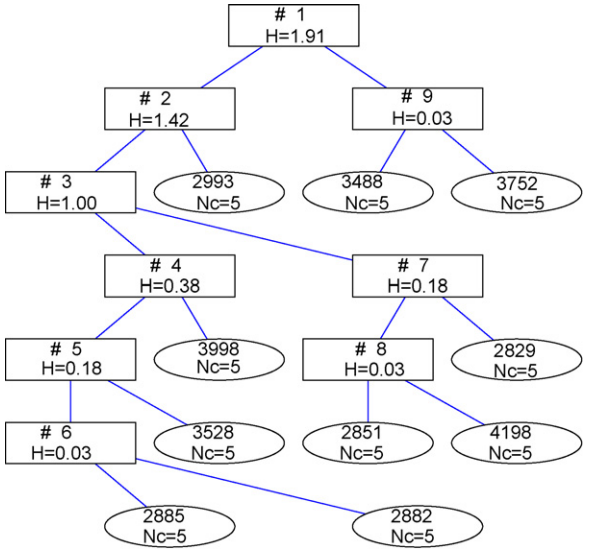

Fig. 6. FuRES classification tree of jet fuels of data collected under the FF-CI mode.

The average confusion matrix from the 10 bootstraps for the FuRES prediction in the FF-CI mode is given in Table 5. The sample classes correspond to the rows and the predicted classes correspond to the columns. The average number and confidence intervals of correct predictions comprise the diagonal elements of the matrix and the off-diagonal elements are the erroneous predictions. In Table 5, sample 4198 in one of the 10 bootstraps was misclassified as sample 3528 by FuRES. Note in Figs. 1 and 2 that these two lots are similar with respect to jet fuel temperature properties, so the single misclassification in the confusion matrix hints that fuels with similar properties are similar in the multivariate GC × MS data space. A matched sample *t*-test was used to compare the FuRES and o-PLS prediction errors in the FF-CI mode with and without wavelet compression. A *t* statistic of −1.0 with a probability of 0.3 was obtained for both with and without wavelet compression, which indicates no significant difference between these two methods at a 95% confidence level.

The two-way ANOVA results supported the results from the *t*-test for which one way was the comparison between the FuRES and o-PLS prediction treatments and the other was the effect of the jet fuel samples. In the FF-CI mode the treatment effect had probabilities of 32% for both with and without wavelet compression, demonstrating FuRES and o-PLS classifiers giving similar results without a significant difference for the classification. The sample

Table 5
Confusion matrix of FuRES classification under the FF-CI mode with wavelet compression.

	2885	3488	3528	2851	2829	4198	2993	3998	2882	3752
2885	5	0	0	0	0	0	0	0	0	0
3488	0	5	0	0	0	0	0	0	0	0
3528	0	0	5	0	0	0	0	0	0	0
2851	0	0	0	5	0	0	0	0	0	0
2829	0	0	0	0	5	0	0	0	0	0
4198	0	0	0.1 ± 0.2	0	0	4.9 ± 0.2	0	0	0	0
2993	0	0	0	0	0	0	5	0	0	0
3998	0	0	0	0	0	0	0	5	0	0
2882	0	0	0	0	0	0	0	0	5	0
3752	0	0	0	0	0	0	0	0	0	5

Table 6
Comparison of prediction errors for data collected 4 days apart with and without wavelet compression and retention time alignment.

Prediction error out of 50 samples	Batch A model predicts batch B		Batch B model predicts batch A	
	FuRES	o-PLS	FuRES	o-PLS
No alignment				
WL compressed	2	3	7	7
Uncompressed	4	3	20	11
Aligned to mean of batch A				
WL compressed	0	4	7	6
Uncompressed	0	1	0	4
Aligned to mean of batch B				
WL compressed	0	3	4	7
Uncompressed	0	0	0	4

Note: WL designates a 2×2 biorthogonal wavelet compression to 1/16th of the data set size. The numbers in the table are prediction errors out for 5 replicates of 10 samples in each data collection batch.

factor was not significant with both probabilities of 48%, indicating that the different samples did not result in different classification rates. Because o-PLS was used as a positively biased control, the above results further validate the FuRES classifiers. The success of FuRES is attributed to the robustness of the method and the high information content of the two-way data objects. (For example for sample 3998 the original dimensions were 5 563 276 with 4124 for GC and 1349 for MS and the data size after compression was 348 478 with 1031 for GC and 338 for MS.) FuRES affords the opportunity to classify jet fuels with a data collection time of about 7 min for each sample from fast GC and fast scan MS measurements.

3.7. Prediction of a novel collection of samples

To validate the robustness of the proposed method, ten new samples were randomly selected from the library of 200 samples. The same set of samples used in the previous study was no longer available for further analyses. The GC/MS data of the ten new fuel lots were collected and analyzed using the FF-CI mode, in two batches that were separated by a span of 4 days. Data from the earlier collection will henceforth be referred to as batch A, and data from the later collection will be referred to as batch B. The dilution procedure and time period separating these two batches of samples added additional sources of variation. For these studies, the procedure of data collection and classification was implemented with only the addition of an unmodified three-way polynomial retention time alignment.

Retention time drift is caused by fluctuations in inlet and outlet pressures, temperature, column degradation, sample size and flow rate during gas chromatographic measurements. When drift happens classification and prediction rates will be deleteriously affected; retention time alignment is necessary to realign the peaks in the chromatograms to a standard. This approach used an unsupervised alignment that adjusted the retention time by a cubic polynomial of each two-way image to the closest correspondence to the average two-way object.

In the previous study, because the data of fuels were collected during the same time period, retention time alignment did not affect the classification rates.

The first batch of new samples (A) was used to build a FuRES and an o-PLS model to predict the latter batch (B) of the new samples. The roles of the two data sets were reversed with the batch B used for model building and batch A used for prediction. The numbers of prediction errors are given in Table 6.

Retention time alignment significantly decreased the number of prediction errors. Wavelet compression improved the prediction accuracies for the unaligned data because the loss in chromatographic resolution can correct drift problems as unaligned peaks begin to overlap with respect to retention time.

Two alignments were evaluated: (1) the calibration set was aligned and then each prediction object was aligned to the mean of the calibration set; (2) the prediction data were aligned than each calibration object was aligned to the mean of the prediction set. When the data sets were aligned to the mean of batch B, the prediction accuracies were improved, which may be a result of more serious drift within batch A. After compression, for the data aligned to batch B the prediction errors increased for all the cases except for one of the FuRES classifications that retained at 100% accuracy.

The three-way alignment as one would expect to improve the predictions of data collected over a significant time period. These results demonstrate that the classification procedure works for different sets of fuels by lot number, the classification models were general for a 4-day period, and independent dilution and data collection did not deleteriously affect the prediction rate as long as retention time drift was corrected. The FuRES classifiers gave better or equivalent prediction accuracies as the positively biased o-PLS classifiers. Perfect prediction was achieved after retention time alignment for both compressed and uncompressed data with the batch A classification models.

4. Conclusions

Rapid classification of jet fuels can be realized by using fast GC-fast QIT MS combined with chemometric methods. The novelty of this method resides in the application of fast MS as the detection method for fast GC for the purpose of three-way data collection to classify jet fuel samples by lot numbers. The pretreatment methods for data such as compression and retention time alignment have proved useful and may in some cases improve classification performance while reducing the computational load for building models. Three Latin partitions and ten bootstraps were used to validate the FuRES model. The FuRES classification with and without wavelet compression achieved $99.8 \pm 0.5\%$ classification accuracy with jet fuels that are similar with respect to composition and property. The classification accuracies of FuRES had no significant difference from those obtained by the positively biased o-PLS control method. FuRES has the benefit of no adjustable parameters for configuration that PLS has in regard to the number of latent variables to be used for the model.

A second study with ten different lot numbers of jet fuel samples was completed successfully without optimization or changing any of the instrumental or data processing procedures and parameters. Two independent sets of data were collected 4 days apart and some retention time drift occurred. Routine polynomial three-way retention time alignment was used without modification. To make efficient use of the data, the earlier and later collections were each used for prediction and model building. Besides the incorporation

of the three-way retention time alignment no other modifications were used to the parameters of the procedures.

For these new data sets, 2×2 wavelet compression deleteriously affected prediction rates. For uncompressed data, FuRES achieved perfect prediction (100%) for four different models. These results demonstrate that the proposed method is robust and validated. With this novel method, analysis time was reduced with respect to conventional GC/MS analysis and prediction accuracy improved with respect to fast gas chromatography with normal scan mass spectrometry.

Acknowledgements

The Center for Intelligent Chemical Instrumentation of Ohio University and the Wright Patterson Air Force Base are thanked for the support. Yao Lu, Weiying Lu, Zhanfeng Xu, Shannon Cook, and Zeland Muccio are thanked for their help.

References

[1] M.P. Gómez-Carracedo, J.M. Andrade, M.A. Calvinõ, D. Prada, E. Fernández, S. Muniategui, Talanta 60 (2003) 1051–1062.
[2] J.E. Woodrow, Energ. Fuel. 17 (2003) 216–224.
[3] P. Doble, M. Sandercock, E.D. Pasquier, P. Petocz, C. Roux, M. Dawson, Forensic Sci. Int. 132 (2003) 26–39.
[4] Z. Wang, C. Yang, B. Hollebone, M. Fingas, Environ. Sci. Technol. 40 (2006) 5636–5646.
[5] B. Tan, J.K. Hardy, R.E. Snavely, Anal. Chim. Acta 422 (2000) 37–46.
[6] D.W. Mikolaitis, C. Segal, A. Chandy, J. Propul. Power 19 (2003) 601–606.
[7] T.W. Lee, V. Jain, S. Kozola, Combust. Flame 125 (2001) 1320–1328.
[8] B.K. Lavine, H. Mayfield, P.R. Kromann, A. Faluque, Anal. Chem. 67 (1995) 3846–3852.
[9] J. Ritter, V.K. Stromquist, H.T. Mayfield, M.V. Henley, B.K. Lavine, Microchem. J. 54 (1996) 59–71.
[10] G. Wang, J. Karnes, C.E. Bunker, M.L. Geng, J. Mol. Struct. 799 (2006) 247–252.
[11] S.D. Gregg, J.L. Campbell, J.W. Fisher, M.G. Bartlett, Biomed. Chromatogr. 21 (2007) 463–472.
[12] L.M. Balster, S. Zabarnick, R.C. Striebich, L.M. Shafer, Z.J. West, Energ. Fuel. 20 (2006) 2564–2571.
[13] D.D. Link, J.P. Baltrus, K.S. Rothenberger, P. Zandhuis, D.K. Minus, R.C. Striebich, Energ. Fuel. 17 (2003) 1292–1302.
[14] J.R. Long, H.T. Mayfield, M.V. Henley, P.R. Kromann, Anal. Chem. 63 (1991) 1256–1261.
[15] J.L. Wang, W.L. Chen, J. Chromatogr. A 927 (2001) 143–154.
[16] S. Macho, M.S. Larrechi, Trends Anal. Chem. 21 (2002) 799–805.
[17] K.E. Kramer, R.E. Morris, S.L. Rose-Pehrsson, J. Cramer, K.J. Johnson, Energ. Fuel. 22 (2008) 523–534.
[18] G.E. Fodor, K.B. Kohl, Energ. Fuel. 7 (1993) 598–601.
[19] M. Blanco, I. Villarroya, Trends Anal. Chem. 21 (2002) 240–250.
[20] M.P. Gómez-Carracedo, J.M. Andrade, D.N. Rutledge, N.M. Faber, Anal. Chim. Acta 585 (2007) 253–265.
[21] N. Pasadakis, S. Sourligas, C. Foteinopoulos, Fuel 85 (2006) 1131–1137.
[22] D.J. Cookson, C.P. Lloyd, B.E. Smith, Energ. Fuel. 2 (1988) 854–860.
[23] D.J. Cookson, B.E. Smith, Energ. Fuel. 4 (1990) 152–156.
[24] M. Bernabei, G. Bocchinfuso, P. Carrozzo, C.D. Angelis, J. Chromatogr. A 871 (2000) 235–241.
[25] D.J. Cookson, C.P. Lloyd, B.E. Smith, Energ. Fuel. 1 (1987) 438–447.
[26] N.E. Watson, M.M. VanWingerden, K.M. Pierce, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1129 (2006) 111–118.
[27] P. Rearden, P.B. Harrington, J.J. Karnes, C.E. Bunker, Anal. Chem. 79 (2007) 1485–1491.
[28] K.J. Johnson, B.J. Prazen, D.C. Young, R.E. Synovec, J. Sep. Sci. 27 (2004) 410–416.
[29] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 72 (2000) 4154–4162.
[30] L. Cao, Dissertation, Athens, OH, 2004, p. 180.
[31] B.K. Alsberg, R. Goodacre, J.J. Rowland, D.B. Kell, Anal. Chim. Acta 348 (1997) 389–407.
[32] P.B. Harrington, Trends Anal. Chem. 25 (2006) 1112–1124.
[33] G. Liu, L. Wang, H. Qu, H. Shen, X. Zhang, S. Zhang, Z. Mi, Fuel 86 (2007) 2551–2559.
[34] Y.C.E. Chao, R.L. Gibson, L.A. Nylander-French, Ann. Occup. Hyg. 49 (2005) 639–645.
[35] P.B. Harrington, J. Chemom. 5 (1991) 467–486.
[36] G.S. Frysinger, R.B. Gaines, J. High Resol. Chromatogr. 22 (5) (1999) 251–255.
[37] K.J. Johnson, R.E. Synovec, Chemom. Intell. Lab. Syst. 60 (2002) 225–237.
[38] J.A. Cramer, R.E. Morris, B. Giordano, S.L. Rose-Pehrsson, Energ. Fuel. 23 (2009) 894–902.
[39] J.A. Cramer, K.E. Kramer, K.J. Johnson, R.E. Morris, S.L. Rose-Pehrsson, Chemom. Intell. Lab. Syst. 92 (2008) 13–21.
[40] R.E. Morris, M.H. Hammond, J.A. Cramer, K.J. Johnson, B.C. Giordano, K.E. Kramer, S.L. Rose-Pehrsson, Energ. Fuel. 23 (2009) 1610–1618.
[41] E. Matisová, M. Dömötörová, J. Chromatogr. A 1000 (2003) 199–221.
[42] L.N. Williamson, M.G. Bartlett, biomed. Chromatogr. 21 (2007) 664–669.
[43] C.G. Yang, M.E. Bier, Anal. Chem. 77 (2005) 1663–1671.
[44] P.B. Harrington, N.E. Vieira, P. Chen, J. Espinoza, J.K. Nien, R. Romero, A.L. Yergey, Chemom. Intell. Lab. Syst. 82 (2006) 283–293.
[45] P.B. Harrington, C. Laurent, D.F. Levinson, P. Levitt, S.P. Markey, Anal. Chim. Acta 599 (2007) 219–231.
[46] Y. Lu, P.B. Harrington, Anal. Chem. 79 (2007) 6752–6759.
[47] Y. Kumooka, Forens. Sci. Int. 189 (2009) 104–110.